ELSEVIER

Journal of Statistical Planning and Inference 83 (2000) 449–473

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Dual group screening

Wesley O. Johnson[a,*], Joseph L. Gastwirth[b,1]

[a] Division of Statistics, University of California, Davis, CA 95616, USA

[b] Department of Statistics, George Washington University, Washington DC 20052, USA

Received 15 August 1997; accepted 17 February 1999

Abstract

We propose a screening procedure that allows for the possibility of catching units that might be missed by current methods, and enables us to estimate the sensitivity of the screening test used and the prevalence in the initially screened population by using only information that is collected during the course of the procedure. This potential does not exist with current methods. The proposed protocol involves pooling units at two stages in order to both screen and to provide quality control. We propose estimators for all parameters of interest and develop appropriate asymptotic inferences. Simple easy-to-implement formulas are obtained. Since units from first stage negative groups are dependent, the mathematics necessary to keep track of statistical information obtained at the second-stage is quite delicate. Monte Carlo simulations indicate that our asymptotic variance formulas are highly accurate while large sample normality depends on how large the sample size is relative to the concentration of the parameters near the boundary of the parameter space. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Asymptotics; Drug screening; HIV screening; Group testing; Prevalence; Sensitivity; Predictive value negative

1. Introduction

The HIV epidemic focused attention on the need for efficient screening of a large number of units of donated blood. Consequently, the group screening idea of Dorfman (1943) has stimulated a new literature (Gastwirth and Hammick, 1989; Chen and Swallow, 1990; Hammick and Gastwirth, 1994; Kline et al., 1989; Litvak et al., 1994; Hughes-Oliver and Swallow, 1994; Gastwirth and Johnson, 1994), and the recent review by Foulkes (1998).

Much recent work has emphasized the very low prevalence setting. However, in other potential areas of application such as screening job applicants or employees in

[*] Corresponding author.

[1] Research supported in part by a grant from NSF.

ifety-sensitive jobs for drug use, the prevalence may exceed 1%. Because employ- ;s who test positive due to lab error may not have legal recourse, it is important to corporate a quality control procedure in addition to an efficient screening program. his paper incorporates the quality control group testing method in Gastwirth and ›hnson (1994) with group screening of the original samples (Emmanuel et al., 1988; ahoon-Young et al., 1989; Chen and Swallow, 1990; Litvak et al., 1994) accom- ishing both goals. Thus our proposed protocol allows for the possibility of grouping imples initially and for re-testing a fraction of the negatives in groups. We refer to ese stages as first and second-stage grouping and to the overall procedure as dual oup screening.

Our procedure is directed at screening programs for characteristics whose prevalence low or moderate ($<10\%$) and it applies to virtually any situation where it is feasible id efficient to do group screening in the first place. We are concerned with situations here a generally imperfect but relatively inexpensive test is used initially and where a :rfect, but relatively expensive, “gold standard” (GS) test is used to confirm positive sults based on the inexpensive test but which is not used to confirm negative results. IV testing for blood that has been donated for transfusion falls into this category, for cample Nusbacher et al. (1986), Gunson and Rawlinson (1988). In this setting, there is) information in the collected data for estimating the sensitivity (proportion of correct)sitive test outcomes) or the prevalence of the characteristic in the screened population 3astwirth et al., 1991; Johnson and Gastwirth, 1991; Gastwirth and Johnson, 1994). e-pooling some of the negatives from the first-stage (FS) enables us to make more :curate inferences about the prevalence, and the predictive values positive and negative)revalence among the positives and one minus the prevalence among the negatives :spectively), as well as the accuracies of the screening test. A change in the prevalence)uld be detected as well as a change in the sensitivity. An added benefit of our protocol that it allows for the possibility of limiting the number of false negatives that might iss through the system undetected, when this is of interest. Our procedure is not iitable if there is no GS test, if the FS test destroys the testing material leaving none)r a second test, or if the added cost of re-pooling and testing is too high relative to ie benefit of making more accurate statistical inferences or catching false negatives om the FS.

Our protocol also applies to programs where a large number of units are to be :reened and where detecting positive units is very important, as in the case of HIV :reening for example. Ratcliffe et al. (1998) demonstrated that known methods of reventing mother to child transmission are cost-effective. They reduce the probability f transmission from about 0.3 to 0.05, and their cost is small ($1000–1500) relative › the estimated cost of treating an HIV-infected child ($400,000). Moreover, even ·hen the child of a treated mother is infected, on average they have an extra year of .IDS-free life. Thus, the use of group screening in antenatal clinics, especially those i areas in Laos and neighboring areas of Thailand where the prevalence of HIV in regnant women ranges from 0.3 to 5% (Loue et al., 1998) should be cost effective. ıcorporating a second stage may be useful in cases where the prevalence is over 1%.

The added complexity and cost of our protocol may not be justified in situations where an occasional false negative is not a serious issue, e.g., gene transfer experiments (Chick, 1996), or when retesting is not possible, especially since previously uninfected individuals may become infected during the testing process (Thompson, 1962). However, our approach is always applicable when mass screening is applied to grouped samples without confirmatory testing on negative test results, and when it is of interest to ascertain the accuracy of the first-stage screening test.

The optimum choice of the size of the groups used in the two stages depends on the prevalence of the trait in the population as well as the accuracy and costs of the tests. The important issue of deciding on an appropriate group size is discussed at length by Hughes-Oliver and Swallow (1994). Alternatively, one can assume that group sizes are determined primarily so that there is minimal loss in terms of sensitivity and specificity (Sherlock et al., 1995). Kline et al. (1989) and Monzon et al. (1991) have argued that pooling in groups of size 10–20 is quite reasonable for HIV testing. Kantanen et al. (1996) noted a dilution effect for both screening and standard confirmatory tests and thus chose pools of size 5. The World Health Organization has recommended that no more than five sera be pooled for screening for HIV and that pooling would not be effective for populations with a prevalence in excess of 2% (World Health Organization: Global programme on AIDS and global blood safety initiative — recommendations for testing for HIV antibody on serum pools. Weekly Epidemiological Recorder, 1991, 44: 326–327).

In Section 4.3 we provide an expected cost formula for our two-stage procedure which can be used to decide on the appropriateness of pooling at either stage for the specific situation and to select group sizes. Second-stage pooling is virtually always appropriate provided the FS screening test is reasonably sensitive since the prevalence of the characteristic at the second stage will be small. It is thus a simple matter of weighing the cost of second-stage grouping with the added benefit of having more precise statistical inferences and the possibility of catching false negatives.

The model and estimators of prevalence and screening accuracy are presented in Section 2. Asymptotic distribution theory is presented in Section 3 and justified in Appendix A. Monte Carlo simulation results and a discussion of the accuracy of the asymptotic results is given in Section 3.4. Illustrations and a discussion of cost effectiveness and efficiency are discussed in Section 4, and final comments and conclusions are given in Section 5.

2. Model and estimators: dual-grouping

Our proposed procedure has two stages. The first stage (FS) modifies the “standard screening protocol”, where every unit is tested, by randomly pooling the units into groups of size $\tilde{k}\geqslant 1$. As errors inevitably occur at the first stage, we add a second stage where negative units from the first stage can be selected, pooled into new groups and re-tested for the purpose of catching units that were missed at the first stage and/or where $\tilde p_j = \binom{\tilde k}{j} \pi^j (1-\pi)^{\tilde k - j}$. We note that

$$\tilde x \equiv \mathrm{vec}(\tilde X') \sim \mathrm{Mult}(\tilde N, \tilde p = \mathrm{vec}(\tilde P')). \tag{4}$$

The elements in the second row of X are unobserved but their sum, x_{n}, is. The unobserved number of first-stage false negative individuals,

$$x_{\mathrm{fn}} \equiv \sum_{j=1}^{\tilde k} j x_{\mathrm{fn}j} \tag{5}$$

will be used in the asymptotic analysis.

2.2. Second-stage screening

At the second-stage, a random sample of units from the $\tilde k x_{\mathrm{n}}$ negatives is selected. Each negative unit has probability f of being selected and randomly placed into a group of size k. The groups are re-screened and those that test positive are given confirmatory tests. Individual units within groups that are confirmed positive may or may not be given confirmatory tests; such tests would be given to determine precisely which units were defective in a confirmed positive group. In the context of HIV screening one would generally identify the individual positives, but in screening employees this might depend on the negotiations that established the screening procedure, especially as some persons are being tested twice. The fraction, f, associated with the $\tilde k x_{\mathrm{n}}$ units to be re-tested depends on the purpose; to make statistical inferences, f may be modest, e.g. 0.2, but for identifying as many missed units as possible, it should be one.

We denote the sensitivity and specificity of the screening tests for groups by η_{g} and θ_{g}, namely

$$\eta_{\mathrm{g}} = \mathrm{pr}(S \mid \text{at least one } D \text{ out of } k), \quad \theta_{\mathrm{g}} = \mathrm{pr}(\bar S \mid \text{no } D\text{'s out of } k).$$

We have again (see discussion just after (2)) assumed that the sensitivity of the screening test does not increase if the group contains more than one D. This assumption is statistically conservative as more D's would be detected at the second stage if it were not true. As we noted earlier, the prevalence at the second stage is very low so very few groups will have one, much less 2 or more, D's.

The procedure reports the following data at the second-stage:

$$\begin{aligned}
&v = \#\text{ FS negative units identified for re-testing},\\
&m = \left[\frac{v}{k}\right] = \#\text{ second-stage groups},\\
&x_{\mathrm{tp}}^{\mathrm{g}} = \#\text{ second-stage true positive groups},\\
&x_{\mathrm{fp}}^{\mathrm{g}} = \#\text{ second-stage false positive groups},\\
&s_{\mathrm{tp}}^{\mathrm{g}} = \#\text{ second-stage true positive individuals in true positive groups}.
\end{aligned} \tag{6}$$

If individuals within second-stage true positive groups are not confirmed, $s_{\mathrm{tp}}^{\mathrm{g}}$ is not observed.

We note that a reasonable surrogate for the "missing" number of FS false negatives, x_{fn}, is $s_{\mathrm{tp}}^{\mathrm{g}}/f\eta_{\mathrm{g}}$. To see this, first suppose $f = \eta_{\mathrm{g}} = 1$, as would be the case if we re-tested all the negatives and with a GS test. In this instance, $s_{\mathrm{tp}}^{\mathrm{g}} = x_{\mathrm{fn}}$. Now suppose that $f = 0.5$. Then clearly $s_{\mathrm{tp}}^{\mathrm{g}}$ is expected to be half of what it would be if all the FS negatives had been retested, so dividing it by f makes the proper adjustment. Finally let $\eta_{\mathrm{g}} = 0.9$ and $f = 1$. Then if $s_{\mathrm{tp}}^{\mathrm{g}} = 9$, we could infer that 9 out of the 10 available D's were detected so that dividing by η_{g} makes the proper adjustment; dividing by $f\eta_{\mathrm{g}}$ takes care of both issues simultaneously.

In the event that confirmatory testing was not performed on individual units in second-stage positive groups, only $x_{\mathrm{tp}}^{\mathrm{g}}$ would be observed rather than $s_{\mathrm{tp}}^{\mathrm{g}}$. However, since the second-stage prevalence will generally be quite low it follows that $s_{\mathrm{tp}}^{\mathrm{g}} = \sum_j j x_{\mathrm{tp}j}^{\mathrm{g}} \doteq \sum_j x_{\mathrm{tp}j}^{\mathrm{g}} = x_{\mathrm{tp}}^{\mathrm{g}}$, where $x_{\mathrm{tp}j}^{\mathrm{g}}$ is the number of second-stage groups with exactly j D's. Thus, $x_{\mathrm{tp}}^{\mathrm{g}}/f\eta_{\mathrm{g}}$ would then serve as a surrogate for x_{fn}. We make this precise in Section 3.3 and Appendix A.

With $\tilde k = 1$, estimators of $(\pi, \eta, \theta, \pi_1, \psi)$ are obtained by substituting $s_{\mathrm{tp}}^{\mathrm{g}}/f\eta_{\mathrm{g}}$ or $x_{\mathrm{tp}}^{\mathrm{g}}/f\eta_{\mathrm{g}}$ for x_{fn}, the unavailable number of first stage false negatives, in (1). With arbitrary $\tilde k$, estimators of the prevalences π and π_1 are obtained by making this substitution as well and, in the latter case, by also realizing that $\tilde k x_{\mathrm{n}}$ is needed in the denominator. Furthermore, $E(x_{\mathrm{tp}}) = N\pi\eta$ and $E(x_{\mathrm{fp}}) = \tilde N (1-\pi)^{\tilde k}(1-\theta)$ due to (4) and the definitions at (2). We thus define

$$\hat\pi = \frac{x_{\mathrm{tp}} + s_{\mathrm{tp}}^{\mathrm{g}}/f\eta_{\mathrm{g}}}{N}, \quad \hat\eta = \frac{x_{\mathrm{tp}}}{N\hat\pi}, \quad \hat\theta = 1 - \frac{x_{\mathrm{fp}}}{\tilde N (1-\hat\pi)^{\tilde k}}, \quad \hat\pi_1 = \frac{s_{\mathrm{tp}}^{\mathrm{g}}}{\tilde k f \eta_{\mathrm{g}} x_{\mathrm{n}}}, \quad \hat\psi = \frac{x_{\mathrm{tp}}}{\tilde k x_{\mathrm{p}}}, \tag{7}$$

if $s_{\mathrm{tp}}^{\mathrm{g}}$ is observed; otherwise, substitute $x_{\mathrm{tp}}^{\mathrm{g}}$ in which case the estimators are defined to be $\hat\pi^*, \hat\eta^*, \hat\theta^*, \hat\pi_1^*$ and $\hat\psi^*$, respectively. These estimators are shown to be asymptotically normal and consistent in Section 3 and Appendix A.

3. Asymptotic distribution theory

In this section we give the asymptotic distribution theory for the estimators given in (7). The derivation of these results is quite delicate, involving a long sequence of conditioning arguments, and is thus mainly relegated to the appendix. We begin here by giving results for the first-stage statistics as $\tilde N \to \infty$. Then, in Appendix A, we condition on the unobserved number of false negative units from the first stage in order to make calculations for the second stage. Asymptotic results for the second stage are obtained conditionally on FS results and as $m \to \infty$. We then link the asymptotics for both stages to obtain the joint limiting distribution of $(x_{\mathrm{tp}}, x_{\mathrm{fp}}, x_{\mathrm{n}}, s_{\mathrm{tp}}^{\mathrm{g}})$ from which we are able to obtain the limiting distributions for our estimators in (7) by the delta method.

While our results, given in Section 3.2, generalize those in Gastwirth and Johnson (1994) for $\tilde{k}=1$, the method of proof outlined above is totally different. This is due to the fact that the second-stage negative individuals are not independent unless $\tilde{k}=1$, and thus the distribution of $s_{\mathrm{tp}}^{\mathrm{g}}$ is no longer straightforward as it was in Gastwirth and Johnson (1994). The dependence arises from the fact that knowing a particular FS negative individual's status regarding the characteristic is informative for the status of other individuals from the same group. Specifically, if D_i denotes that individual i is D, $\mathrm{pr}(D_2|D_1$, same FS negative group$)=\pi$ which will generally be greater than π_1. We thus condition on x_{fn} in order to account for this dependence. Furthermore, when $f<1$, it will also be necessary to condition on the actual number of D's among the x_{fn} that are re-tested. All such details are in given in Appendix A.

3.1. Asymptotics for the first stage

We require the asymptotic distribution of the vector $F_{\tilde{N}} \equiv (x_{\mathrm{tp}}, x_{\mathrm{fp}}, x_{\mathrm{n}}, x_{\mathrm{fn}})'$. As x_{fn} is unobserved, we *ultimately* integrate over it to obtain the joint distribution of $(x_{\mathrm{tp}}, x_{\mathrm{fp}}, x_{\mathrm{n}}, s_{\mathrm{tp}}^{\mathrm{g}})$. Recall from (4) that $\tilde{x}=\mathrm{vec}(X')\sim \mathrm{Multinomial}(\tilde{N},\tilde{p})$. It follows that $\sqrt{\tilde{N}}(\tilde{x}/\tilde{N}-\tilde{p}) \xrightarrow{\mathrm{L}} \mathrm{N}(0,\Sigma_{\tilde{x}})$ where $\Sigma_{\tilde{x}}=\mathrm{Diag}\{\tilde{p}\}-\tilde{p}\tilde{p}'$. Define the matrix

$$A=\begin{pmatrix} 0 & \tilde{a}' & 0 & 0e_{\tilde{k}}' \\ 1 & 0e_{\tilde{k}}' & 0 & 0e_{\tilde{k}}' \\ 0 & 0e_{\tilde{k}}' & 1 & e_{\tilde{k}}' \\ 0 & 0e_{\tilde{k}}' & 0 & \tilde{a}' \end{pmatrix}$$

with $\tilde{a}=(1,2,\ldots,\tilde{k})'$ and $e_{\tilde{k}}$ a vector of ones. Note that $A\tilde{x}=F_{\tilde{N}}$, and $A\tilde{p}=(\tilde{k}\pi\eta, \tilde{p}_0(1-\theta), q, \tilde{k}\pi(1-\eta))' \equiv \mu_{\tilde{k}}$. Define $\mu=(\mu_1,\mu_2,\mu_3,\mu_4)'=(\pi\eta, \tilde{p}_0(1-\theta), q, \pi(1-\eta))'$. Then we have

$$F_{\tilde{N}}^{*} \equiv A\sqrt{\tilde{N}}(\tilde{x}/\tilde{N}-\tilde{p})=\sqrt{\tilde{N}}(F_{\tilde{N}}/\tilde{N}-\mu_{\tilde{k}}) \xrightarrow{\mathrm{L}} F^{*} \sim \mathrm{N}(0,\Sigma_F), \quad \Sigma_F \equiv A\Sigma_{\tilde{x}}A'). \tag{8}$$

After some algebra,

$$\Sigma_F=\begin{pmatrix} \tilde{k}\mu_1\{1-\tilde{k}\mu_1+\pi(\tilde{k}-1)\} & -\tilde{k}\mu_1\mu_2 & -\tilde{k}\mu_1\mu_3 & -\tilde{k}^2\mu_1\mu_4 \\ * & \mu_2\{1-\mu_2\} & -\mu_2\mu_3 & -\tilde{k}\mu_2\mu_4 \\ * & * & \mu_3(1-\mu_3) & \tilde{k}\mu_4(1-\mu_3) \\ * & * & * & \tilde{k}\mu_4\{1-\tilde{k}\mu_4+\pi(\tilde{k}-1)\} \end{pmatrix}. \tag{9}$$

3.2. Asymptotic distributions for $\hat{\pi}$, $\hat{\eta}$, $\hat{\theta}$, $\hat{\pi}_1$, and $\hat{\psi}$

It is shown in Appendix A that, with specified η_{g},

$$\sqrt{\tilde{N}}\begin{pmatrix} \dfrac{x_{\mathrm{tp}}}{\tilde{N}} & - & \tilde{k}\mu_1 \\ \dfrac{x_{\mathrm{fp}}}{\tilde{N}} & - & \mu_2 \\ \dfrac{x_{\mathrm{n}}}{\tilde{N}} & - & \mu_3 \\ \dfrac{s_{\mathrm{tp}}^{\mathrm{g}}}{\tilde{N}} & - & \tilde{k}\mu_7 \end{pmatrix} \xrightarrow{\mathrm{L}} \mathrm{N}_4(0,\Sigma), \tag{10}$$

where $\mu_7=f\eta_{\mathrm{g}}\mu_4$ and

$$\Sigma=\begin{pmatrix} \tilde{k}\mu_1\{1-\tilde{k}\mu_1+\pi(\tilde{k}-1)\} & -\tilde{k}\mu_1\mu_2 & -\tilde{k}\mu_1\mu_3 & -\tilde{k}^2\mu_1\mu_7 \\ * & \mu_2(1-\mu_2) & -\mu_2\mu_3 & -\tilde{k}\mu_2\mu_7 \\ * & * & \mu_3(1-\mu_3) & \tilde{k}(1-\mu_3)\mu_7 \\ * & * & * & \sigma_{77}^2 \end{pmatrix}$$

and where

$$\sigma_{77}^2=\tilde{k}\mu_7(1-\tilde{k}\mu_7)+\tilde{k}\mu_7^2\left\{\frac{k-1}{f\mu_3}\left(\frac{1}{\eta_{\mathrm{g}}}-1\right)+(\tilde{k}-1)\frac{1}{1-\eta}\right\}.$$

Note that $\Sigma=\Sigma_F$ when $f=\eta_{\mathrm{g}}=1$, which is of course as it should be.

From (10), we obtain the marginal asymptotic normal distributions for the estimators defined at (7). These are all obtained by using the delta method and they involve tedious but simple algebra, so we simply catalogue the results. For each of the five estimators considered, we obtain results just like $\sqrt{\tilde{N}}(\hat{\pi}-\pi) \xrightarrow{\mathrm{L}} \mathrm{N}(0,\sigma_\pi^2)$, for example. All that is needed then are the asymptotic variances and we are done. We obtain

$$\begin{aligned} \sigma_\pi^2 &= \frac{\pi}{\tilde{k}}\left\{1-\pi+(1-\eta)\left(\frac{1}{f\eta_{\mathrm{g}}}-1\right)+\pi(1-\eta)^2q^*\right\}, \\ \sigma_\eta^2 &= \frac{\eta(1-\eta)}{\tilde{k}\pi}\left\{1+\eta\left(\frac{1}{f\eta_{\mathrm{g}}}-1\right)+\pi\eta(1-\eta)q^*+\pi(\tilde{k}-1)\right\}, \\ \sigma_\theta^2 &= \frac{\theta(1-\theta)}{1-\pi}+\frac{\tilde{k}(1-\theta)^2\pi(1-\eta)}{(1-\pi)^2}\tilde{q} \\ &\quad +\frac{1-\theta}{1-\pi}\left\{\frac{1}{(1-\pi)^{\tilde{k}-1}}-1-(1-\theta)(\pi(\tilde{k}-1))\right\}, \\ \sigma_{\pi_1}^2 &= \frac{\pi_1(1-\pi_1)}{\tilde{k}q}+\frac{\pi_1}{\tilde{k}q}\left(\frac{1}{f\eta_{\mathrm{g}}}-1\right)+\frac{\pi_1^2}{\tilde{k}}\left\{(\tilde{k}-1)\left(\frac{1}{1-\eta}-\frac{1}{q}\right)+q^*\right\}, \\ \sigma_\psi^2 &= \frac{\psi(1-\psi)}{\tilde{k}(1-q)}+\psi^2\left(\frac{\tilde{k}-1}{\tilde{k}\eta}\right), \end{aligned} \tag{11}$$

where

$$\bar{q}=\left\{\frac{1}{f\eta_g}-1\right\}+q^*\pi(1-\eta),\quad q^*=\frac{k-1}{f\mu_3}\left(\frac{1}{\eta_g}-1\right).$$

These formulas reduce to those expected if $f=\eta_g=k=\tilde{k}=1$. Note that the right-hand term in brackets in the expression for σ_θ^2 is non-negative since it is increasing in π for fixed θ with minimum at $\pi=0$. The last term in brackets in the expression for $\sigma_{\pi_1}^2$ is also non-negative provided $1-\eta\leqslant\theta$ since this implies that $1-\eta\leqslant q$; this will be true in any reasonable setting. The first three formulas coincide with those obtained in Gastwirth and Johnson (1994) with $\tilde{k}=1$, and the formula for $\sigma_{\pi_1}^2$ coincides, up to $O(\pi_1^2)$, with the formula in GJ for the small π_1 case they consider. Finally, note how the variances vary as a result of group size, accuracy of tests, and the fraction, f. There is no variance inflation for FS grouping when estimating π since the asymptotic variance is free of $\tilde{k}$. The asymptotic variance for $\hat{\psi}$ is free of k since the estimator is not a function of s_{tp}^g. A larger FS group results in a larger asymptotic variance for $\hat{\psi}$, etc. Also note the variance inflation for $f\eta_g<1$.

The above presumes that the correct value of η_g has been specified. In practice η_g will not be known unless a "gold standard" is used at the second stage. Therefore, it is generally necessary to obtain an independent estimate.

We assume an unbiased estimate is available, say $n\hat{\eta}_g\sim\text{Bin}(n,\eta_g)$, perhaps based on a previous study. Our estimates at (7) are revised so that $\hat{\eta}_g$ is substituted for η_g in all formulas. We also assume that $n/\tilde{N}\to c$ as $\tilde{N}\to\infty$. Then the asymptotic variances of the corresponding normalized estimates are identical to the terms obtained in (11) plus an additional term which is due to the uncertainty in the estimate of η_g. These terms are

$$\frac{\{\pi(1-\eta)\}^2(1-\eta_g)}{c\eta_g},\ \frac{\{\eta(1-\eta)\}^2(1-\eta_g)}{c\eta_g},\ \frac{\{\tilde{k}\pi(1-\theta)(1-\eta)\}^2(1-\eta_g)}{c(1-\pi)^2\eta_g},$$

$$\frac{\pi_1^2(1-\eta_g)}{c\eta_g},\ 0$$

for π,η,θ,π_1 and ψ, respectively. Note that these terms will generally be relatively small when the second-stage prevalence is small; the term for η will be relatively small if the two sensitivities are large.

3.3. No confirmation of second-stage individual units

We rely on the asymptotic conditional distribution results (23) and (24) in Appendix A to justify substituting x_{tp}^g for s_{tp}^g when π_1 is small and s_{tp}^g is unobserved and x_{tp}^g is. Simple calculations establish that the ratio of the mean (variance) in (24) to the mean (variance) in (23) is $1+O(\pi_1)$, and thus if π_1 is small enough, these moments will be virtually identical. It follows that results established for s_{tp}^g in Appendix A will be identical to those with x_{tp}^g substituted for s_{tp}^g, up to the $O(\pi_1)$ difference in

mean and variance terms. Thus, when the second-stage prevalence is small, which it will be in general, we may substitute x_{tp}^g for s_{tp}^g.

3.4. Monte Carlo simulation

Here, we briefly consider the quality of the asymptotic results. A computer program was written to simulate dual-screening data according to our protocol. Through repeated Monte Carlo (MC) sampling (MC sample size = 1000), we obtained MC approximations to the mean and standard deviations (s.d.) of our estimators (7), and to the confidence level associated with nominal 95% intervals, with η_g "known". We considered situations with $\tilde{N}$ ranging from 50 to 3127, π ranging from 0.001 to 0.2, first-stage accuracies ranging from 0.8 to 0.99, second-stage accuracies ranging from 0.7 to 0.9 with $\hat{\eta}_g$ within 0.1 of η_g, k and $\tilde{k}$ ranging from 5 to 10, and with $f=1$. For all of the situations considered, we found the asymptotic variance formulas (11) to work remarkably well. Estimators (7) were found to be unbiased with the exception of a few situations where $\hat{\eta}_g\neq\eta_g$, and in those cases, the bias was not particularly large. As an illustration of a situation with the largest biases we observed, consider $\tilde{N}=50$, $\pi=0.05$, $\eta=\theta=0.8, \eta_g=\theta_g=0.7$, $\hat{\eta}_g=0.8$, $\tilde{k}=k=10$. Here, we have approximate biases of 0.001, 0.022 and 0.002 for estimators of π,η and π_1 respectively, and we have corresponding s.d.'s 0.010, 0.104 and 0.010 based on (11), and their MC counterparts 0.0099, 0.100 and 0.0095, respectively.

While we found that our formulas (11) worked very well over the above range of possibilities, it was not necessarily the case that the sampling distributions of our statistics (7) had a nice "bell" shape. For example, in the specific instance referred to above, the histograms for the simulated $\hat{\pi}$'s and $\hat{\psi}$'s were bell-shaped, but those for $\hat{\eta}$, $\hat{\theta}$ and $\hat{\pi}_1$ were noticeably skewed and coverage levels for nominal 95% intervals for π,η and π_1 were 0.95, 0.91 and 0.90, respectively. However, there is a trend that we found throughout our study. If the mean plus or minus three times the standard deviation excluded zero and one, all histograms looked reasonably bell-shaped albeit some might have been slightly skewed, and confidence interval levels were reasonably close to 0.95. The amount of skewness was observed to diminish for situations with means that were more distant, in terms of the number of standard deviations, from 0 or 1. Note that this rule works for the above illustration.

How large the sample size must be in order for our asymptotics to apply depends on the magnitudes of the parameters we are trying to estimate. The closer they are to zero or one, the larger the sample size must be. To illustrate, we discuss several more scenarios. First, we attempted situations like the one already discussed. It was not until we increased π to 0.2, and decreased all accuracies to 0.7 that histograms looked reasonably bell shaped; confidence levels for π,η and π_1 were 0.95, 0.93, and 0.94. Then we considered $\tilde{N}=100$ with $\pi=0.05,\eta=\theta=0.95,\eta_g=\theta_g=0.9,\hat{\eta}_g=0.8$. In this instance, the estimators are nearly unbiased with MC s.d.'s very close to their asymptotic counterparts based on (11). The s.d. for π is 0.0069 so, since zero is over 8 s.d.'s less than 0.05, we expect the histogram for $\hat{\pi}$ to look bell shaped, which it does.

However, the s.d. for $\hat{\eta}$ is 0.04 and the corresponding histogram is decidedly skewed, as is the one for $\hat{\pi}_1$ since $\pi_1 = 0.0042$ and the corresponding s.d. is 0.0035. Monte Carlo coverages were 0.95, 0.83 and 0.84, respectively. Increasing π to 0.2 and decreasing all accuracies to 0.8 results in reasonable looking histograms with MC coverages of 0.95, 0.92 and 0.93. With $\tilde{N} = 500$, $\pi = 0.05$, $\eta = \theta = 0.95$, $\eta_g = \theta_g = 0.9$, $\hat{\eta}_g = 0.8$, there was no bias for π, η, or π_1, corresponding MC standard deviations were virtually identical to those based on (11), and MC coverages were 0.96, 0.93 and 0.94. With $\tilde{N} = 1000$, $\pi = 0.02$ and the same values as just above, results were again good, and when π is increased to 0.05, they were very good; MC coverages were 0.95, 0.93 and 0.94.

Based on the simulation it appears that the estimates and the formulas for their standard deviations are appropriate in the settings for which our protocol is appropriate. For the large sample normal theory to apply, the simple check of whether the interval defined by the estimate plus or minus 3 standard deviations does not contain 0 or 1 suffices. When this is the case, the 95% confidence intervals derived from the asymptotic results will have coverage near 0.95. If higher levels of confidence are required, then one should replace three standard errors by four or five in the criteria.

4. Illustrations and cost/efficiency comparisons

Since our proposal for dual group screening is new, we present two illustrations based on real data which have been augmented to fit our situation. We then assess the number of defective units that will be detected at the second stage. And finally, we demonstrate how to determine when dual group screening is cost effective and relatively efficient when prevalence, accuracies, and the costs of screening and confirmatory tests are assumed known.

4.1. Illustrations

4.1.1. Drug Testing

Smith–Kline, a major drug testing firm, reported that 3.1% of transportation workers tested positive for drug use (Chemical Regulation Reporter, 1990, p. 781). Gastwirth and Johnson (1994, p. 975) constructed an artificial data set that was consistent with known properties of drug tests and which resulted in the above estimate of prevalence for their dual screening procedure. Here, we modify that data set so that it could have resulted from dual group screening. We first set $\tilde{k} = 10$ and $k = 20$ and assume an independent estimate of η_g is available in the form of a binomial $(n = 100, \eta_g)$ random variable. We assume $\hat{\eta}_g = 0.7$ was observed. Assume that the $N = 3,200,000$ transportation workers are subject to drug testing and that the following data were collected: $x_{\text{tp}} = 93,000$, $x_{\text{fp}} = 10,000$, and $x_{\text{n}} = 217,000$. Table 1 gives our estimates, their standard errors (se) and asymptotic 95% confidence intervals (CI) for f=0.2, 0.05, and 0.01 for the given values in the table. Precision for π, θ, and ψ is clearly excellent for f as small as 0.01. Precision for η and π_1 is good to very good for f as small as 0.05. With $f = 1$, all inferences would be extremely precise due to the very large sample size. Precision would be improved with larger η_g and larger N. Note that in all instances, estimates are many standard deviations above 0 and below 1 so our asymptotic results should be quite adequate in all respects.

Table 1
Artificial drug data: Estimates, standard errors and asymptotic 95% CIs with $f = 0.2$, $s^{\text{g}}_{\text{tp}} = 869$; se's with $f = 0.05$, $s^{\text{g}}_{\text{tp}} = 869/4$, and $f = 0.01$, $s^{\text{g}}_{\text{tp}} = 869/20$

Parameter	$f = 0.2$			$f = 0.05$	$f = 0.01$
	Estimate	se	CI	se	se
π	0.0310	0.00012	(0.0308, 0.0312)	0.00016	0.00031
η	0.9374	0.00456	(0.9287, 0.9461)	0.00557	0.00977
θ	0.9572	0.00042	(0.9563, 0.9580)	0.000427	0.000443
π_1	0.00286	0.0000995	(0.00266, 0.00306)	0.000196	0.000438
ψ	0.0482	0.000227	(0.0477, 0.0486)	0.000227	0.000227

Table 2
Artificial HIV data: Estimates, *se*'s and asymptotic 95% CIs with $s^{\text{g}}_{\text{tp}} = 4f$

Parameter	$f = 1$			$f = 0.5$	$f = 0.25$
	Estimate	se	CI	se	se
π	0.0090	0.00053	(0.0079, 0.0100)	0.00054	0.00090
η	0.9850	0.0077	(0.9695, 1[a])	0.0107	0.0150
θ	0.9843	0.0023	(0.9796, 0.9889)	0.0023	0.0023
π_1	0.00176	0.00091	(0[a], 0.00358)	0.00126	0.00178
ψ	0.0380	0.0013	(0.035, 0.041)	0.0013	0.0013

[a] Truncated at 0 or 1.

4.1.2. HIV Testing

We consider data which was collected by Sherlock et al. (1995) for the purpose of determining seroprevalence of HIV in British Columbia. For illustration, we focus on their data for men. Out of 31,271 men tested in groups of size ten, 276 individuals were ultimately determined to be infected. We thus assume that $\tilde{N} = 3127$ groups of size $\tilde{k} = 10$ were formed and that $x_{\text{tp}} = 276$ true positive individuals were observed. As this is all that can be ascertained with certainty, we proceed to formulate an imaginary data set that is consistent with this one. In the larger data set presented in Sherlock et al. (1995), there were nine false positive groups out of 600 that screened positive, and thus we set $x_{\text{fp}} = 45$ to be roughly consistent with that information. Assume $x_{\text{n}} = 2,390$. We let $k = 10$ and assume $s^{\text{g}}_{\text{tp}} = 4f$ is observed for $f \in \{1, 0.5, 0.25\}$. Table 2 gives inferences for the basic five parameters. The precision is quite good. There is again little or no loss in accuracy for π, θ, and ψ with smaller values of f while the *se*'s for η

and π_1 are increased by a factor of 2 in reducing f from 1 to 0.25. Note that while our asymptotic standard errors will be accurate in this setting, the large sample normality of the results is questionable for η and π_1 since the corresponding CIs overlap one and zero, respectively.

As part of our MC study, we simulated data with parameters that were very similar to the estimates obtained in Table 2. In that simulation, estimates were very close to unbiased, MC standard deviations were exceptionally close to those based on (11), and MC coverage levels for π, η and π_1 were 0.96, 0.91 and 0.91, respectively. If $fs^{\mathrm{g}}_{\mathrm{tp}}$ had been observed to be 10 or larger, the asymptotic normality would not have been called into question according to our suggested criterion (estimate plus or minus 3 standard errors not overlapping zero or one).

4.2. Catching missed units

We consider the potential benefit of second stage screening in terms of detecting units that would otherwise have been missed. The number of first-stage false negative units that are detected at the second stage is $s^{\mathrm{g}}_{\mathrm{tp}}$. With FS individual testing and no second stage, the expected number of detected units is $N\pi\eta$, while with dual group testing it is approximately $N\pi\eta + Nf\pi(1-\eta)\eta_{\mathrm{g}}$, regardless of $\tilde{k}$ and k, and provided the FS group test has the same sensitivity as a single test. For the conditions listed in Table 3 with $f = 1$, $N = 65{,}430$, and $\pi = 0.01$, say, we expect to catch around 647 of the 654 expected defective units with individual sampling, while we would expect to catch virtually all of the defective units with the dual-grouping procedure.

The higher the prevalence, the greater the potential for dual screening in terms of catching appreciable numbers of individuals that would otherwise have escaped detection. As the prevalence of characteristics like HIV-infection varies widely one could save money by instituting a group testing quality control system in regions of relatively high prevalence, e.g., greater than 0.001. If further studies corroborate that the accuracy of the current screening tests are reasonably high for groups of size 5 and 10, the savings in lives alone with dual testing would justify its use. Further research is needed to determine the best combination of the first-stage method and the quality control (second-stage) screen.

Table 3
$M = \$10^6,\ f = 1,\ \eta = 0.99,\ \theta = 0.995,\ \eta_{\mathrm{g}} = 0.98,\ \theta_{\mathrm{g}} = 0.99,\ k = \tilde{k} = 10,\ c_{\mathrm{S}} = \$2.50 = c_{\tilde{k}}$

π	$c_{ys} = \$5$				$c_{gs} = \$25$			
	N_1	N_1/N_2	$\mathrm{sd}_1(\hat{\eta})$	$\mathrm{var}_2(\hat{\eta})/\mathrm{var}_1(\hat{\eta})$	N_1	N_1/N_2	$\mathrm{sd}_1(\hat{\eta})$	$\mathrm{var}_2(\hat{\eta})/\mathrm{var}_1(\hat{\eta})$
0.0001	990,100	3.0	0.010	3.0	941,800	2.97	0.010	2.96
0.001	948,100	2.87	0.0033	2.85	767,300	2.44	0.0037	2.41
0.01	674,200	2.07	0.0013	1.91	276,700	0.95	0.002	0.87
0.05	331,600	1.09	0.00094	0.75	84,200	0.38	0.0019	0.26

4.3. Cost/efficiency considerations

As previously mentioned, it will not be cost effective to group individual units in the first stage if the prevalence is above 0.1 since too many FS groups would test positive. However, grouping at the second stage is feasible since the second-stage prevalence will generally be low. In this subsection, we give some formulas for determining whether or not dual grouping is effective for particular situations.

The expense associated with screening tests derives from the cost of the test kits which are used and from the time and effort on the part of laboratory technicians. Thus the costs differ due to the difference in technician time. Let the cost of a screening kit be c_{S} for both stages. Assume the cost of technician time for a group of size r is c_r for $r = 1, 2 \ldots$. Let the cost of a confirmatory test be c_{gs}. Then for a given sample size, N, and fixed group sizes, the expected cost of first-stage grouping is

$$E(\mathrm{Cost}\,1) \equiv \tilde{N}[(1-q)c_{\mathrm{gs}} + \tilde{k}\chi(\tilde{k} > 1)\{1-(1-\pi)^{\tilde{k}}\}\eta c_{\mathrm{gs}} + c_{\mathrm{S}} + c_{\tilde{k}}], \tag{12}$$

where $\chi(\cdot)$ denotes the indicator function. The approximate expected cost of second-stage grouping is

$$\begin{aligned} E(\mathrm{Cost}\,2) \equiv \frac{Nqf}{k}[\{(k+1)\{1-(1-\tilde{\pi}_1)^k\}\eta_{\mathrm{g}}c_{\mathrm{gs}} \\ +(1-\tilde{\pi}_1)^k(1-\theta_{\mathrm{g}})c_{\mathrm{gs}} + c_{\mathrm{S}} + c_k\}]. \end{aligned} \tag{13}$$

In the event that only $x^{\mathrm{g}}_{\mathrm{tp}}$, rather than $s^{\mathrm{g}}_{\mathrm{tp}}$, is observed at the second stage, we obtain

$$E(\mathrm{Cost}\,2^*) \equiv \frac{Nqf}{k}[\{\{1-(1-\tilde{\pi}_1)^k\}\eta_{\mathrm{g}}c_{\mathrm{gs}} + (1-\tilde{\pi}_1)^k(1-\theta_{\mathrm{g}})c_{\mathrm{gs}} + c_{\mathrm{S}} + c_k\}]. \tag{14}$$

The total expected cost, $E(\mathrm{Cost})$, is the sum.

We first consider a brief illustration of how one can use our formulas to decide whether or not a second-stage is warranted. Consider a situation where group screening is being done and it is of interest to decide whether or not to use a second-stage screen as well. Suppose 100,000 units are to be screened and that reasonable guesses for FS sensitivity and specificity are 0.95 and 0.99, respectively, and that second-stage accuracies are 0.9, and that $\tilde{k} = 10$, $k = 50$. Assume the prevalence is 0.02, the cost of a GS is $c_{\mathrm{gs}} = \$25$, the cost of a screening test is $c_{\mathrm{S}} = \$2.5$, that the cost of pooling per unit pooled is \$0.25, and that costs are the same for FS and second-stage pooling. Finally, assume that x_{tp} is observed in the FS, but only $x^{\mathrm{g}}_{\mathrm{tp}}$ is observed at the second-stage. Then with $f = 0.5$, the expected cost for FS sampling is \$241,976 (using formula (12)) while the expected cost for second-stage sampling is \$13,477 (using formula (14)), which is about 5% of the total cost. Re-testing individuals in confirmed positive second-stage groups results in a second-stage cost of \$35,316 (using formula (13)). The standard deviation of $\hat{\eta}$ is 0.0101 which can be compared with an "optimal" s.d. computed with the use of a GS test at the second stage and with $f = 1$ of 0.0049. Increasing the prevalence to 0.05 and reducing f to 0.25 results in FS cost of \$499,919 and second-stage cost of $5,356, which is around 1% of the total. Confirming individuals in true positive second-stage groups results in a second-stage cost of $30,845 (using formula (13)). The s.d. for $\hat{\eta}$ is 0.0031 compared with the "optimal" value of 0.0010.

The protocol with $\tilde{k}=1$ will be preferable to dual group screening if the prevalence of the trait in the screened population is too high. The cutoff will depend on the relative costs. In order to illustrate how to select $\tilde{k}$, imagine that a fixed amount of money, say M, is available for screening and that the objectives are to screen as many individuals as possible and to obtain accurate estimates of the parameters. Then let N_1 denote the sample size which results in an $E(\text{Cost})=M$, for dual grouped screening. Next, let N_2 denote the corresponding sample size with $\tilde{k}=1$. The N_i's will vary as we change the scenario. Larger sample sizes for the same cost would generally be preferable. Suppose that there is particular interest in estimating the sensitivity of the FS screening test. Assume that $\tilde{k}$ has been selected to be small enough that the sensitivity for testing FS groups is the same as that for testing individuals, in which case it is appropriate to compare variances of estimators for the two types of dual screening. Let $\text{var}_i(\hat{\eta})$: $i=1,2$, denote the variance of the corresponding estimators of η based on FS grouping ($i=1$) and FS individual testing ($i=2$). These formulas are obtained from (11).

Now consider a scenario which is comparable to the HIV illustration discussed in the previous subsection. Table 3 gives values of N_1, N_1/N_2, $\text{sd}_1(\hat{\eta})$ and $\text{var}_2(\hat{\eta})/\text{var}_1(\hat{\eta})$ for the conditions listed there. If we were using a "gold standard" which costs $5 to administer, we would prefer to use dual group screening at least for $\pi \leqslant 0.01$, and possibly even for $\pi \leqslant 0.05$. While we are able to screen about 10% more units if $\pi = 0.05$, the variance of $\hat{\eta}$ is 33% higher under dual grouping. On the other hand, if $c_{\text{gs}} = \$25$ under the same circumstances, there is a clear preference for $\tilde{k}=1$ if $\pi = 0.05$, a slight preference for $\tilde{k}=1$ if $\pi = 0.01$, and a clear preference for dual group screening if $\pi \leqslant 0.001$. It is evident that with larger costs for the "gold standard" relative to the cost of regular screening, smaller prevalences are required to prefer dual grouping. This is due to the fact that, as this relative cost increases, it is increasingly expensive to re-test all individuals in true positive groups and the probability of a true positive group increases with larger prevalences.

As an alternative strategy, one could re-test individuals in positive groups with a less expensive screening test and then confirm only positive units with the gold standard cf. Kantanen et al. (1996). Modifying the expected cost formulae appropriately, the analogous results to those in Table 3 suggest that dual grouping would be comparable to or preferable to $\tilde{k}=1$ for all circumstances in Table 3. Additional comparisons with $k=\tilde{k}=1$, denoted as case 3, result in ratios for $\text{var}_3(\hat{\eta})/\text{var}_1(\hat{\eta})$ in the range from 1.2 to 5 with $c_{\text{gs}} = \$5$ and from 0.38 to 5 with $c_{\text{gs}} = \$25$.

A more extensive analysis could be performed where the variances of estimators of other parameters were considered. Recall that the asymptotic variances of $\hat{\pi}$ are free from $\tilde{k}$ so that grouping or not at the first stage is irrelevant for estimating π. See Gastwirth and Johnson (1994) for further discussion of the cost-effectiveness of dual screening with $\tilde{k}=1$ when compared with the current procedure of FS individual testing only.

5. Conclusions

We have introduced a new method of collecting and analyzing screening data whe the main goal can be either to make inferences about the population prevalences (b fore and after FS screening) and FS screening test accuracies, or to remove certa individuals from a given population, or both. Simple and accurate hand calculator fo mulas are provided, which can be used to generate confidence intervals or hypothes tests about the various parameters when the sample size is large enough. In particula one can easily use our results to check whether the accuracy of the FS screening te has declined over time. This is very useful when the detection of positive units of primary importance. Pooling samples limits costs and dual-screening provides f quality control.

Acknowledgements

The authors thank Professor J. Sethuraman and two referees for helpful commen and suggestions.

Appendix A.

A.1. Asymptotics for the second stage

All of our distribution theory for this section will be conditional on $F^*_{\tilde{N}}$, define in (8), which contains the FS information. We assume the sequence of condition values tends to a vector of numbers F^* in the range of possible values for the limitin distribution.

As $\tilde{N} \to \infty$, the random variable $v \to \infty$ in probability since $v \sim \text{Bin}(\tilde{k}x_{\text{n}}, f)$ an $x_{\text{n}} \to \infty$ in probability as $\tilde{N} \to \infty$. Thus, $m = [v/k] \to \infty$ in probability. It will necessary to condition on the actual number of D's among the x_{fn} that are re-tested. S define D' to be the number of D's out of the x_{fn} that make it through the first screen a are selected for pooling and re-testing. We proceed to obtain the asymptotic condition distribution of $m, D'|F_{\tilde{N}}$, properly normalized.

We first note that

$$km|_{F_{\tilde{N}}} \sim km|_{x_{\text{n}}} \sim \text{Bin}(\tilde{k}x_{\text{n}}, f),$$

$$D'|_{F_{\tilde{N}},m} \sim D'|_{x_{\text{n}},x_{\text{fn}},m} \sim \text{Hypergeometric}(\tilde{k}x_{\text{n}}, x_{\text{fn}}, km).$$

Since $x_{\text{n}} \xrightarrow{\text{P}} \infty$ as $\tilde{N} \to \infty$, we have

$$\sqrt{\tilde{k}x_{\text{n}}}\left(\frac{mk}{\tilde{k}x_{\text{n}}} - f\right)\Bigg|_{F^*_{\tilde{N}}} \xrightarrow{\text{L}} \text{N}(0, f(1-f)),$$

from which we obtain

$$V_{\tilde N}^{*} \equiv \sqrt{\tilde N}\left(\frac{mk}{\tilde k x_{\mathrm{n}}} - f\right)\Bigg|_{F_{\tilde N}^{*}} \xrightarrow{\mathrm{L}} V^{*} \sim \mathrm{N}\left(0, \frac{f(1-f)}{\tilde k q}\right) \tag{15}$$

by Slutsky's theorem. More elaborate versions of this same type of argument are employed repeatedly below. The results (8) and(15) can also be obtained directly from local limit theorem (Okamoto, 1959) arguments, which implies a type of "strong convergence" which we discuss below.

We require a justification that the limiting joint distribution for $(F_{\tilde N}^{*}, V_{\tilde N}^{*})$ is the distribution one obtains from combining the limiting conditional for $V^{*}|F^{*}$ and the limiting marginal for F^{*}. In this event, since the limiting conditional in (15) is free from F^{*}, we obtain that $V_{\tilde N}^{*}$ is asymptotically independent of $F_{\tilde N}^{*}$.

Okamoto (1959) essentially shows that his continuous approximation, which applies directly to discrete distributions like ours, converges in accord with Sethuraman's (1961) definition of strong convergence. This strong convergence applies to both our conditional and marginal results, and consequently Sethuraman's Theorem 1 applies directly to the continuous versions. Thus, we obtain strong joint asymptotic convergence of the versions to the prescribed joint distributions, which implies convergence in distribution. Okamoto (1959) further shows that the CDFs for the discrete random variables and their corresponding continuous versions become arbitrarily close for large n. So, while our actual distributions do not converge strongly, they are arbitrarily close, on intervals, to distributions that do converge strongly. All of our convergence results below can be formulated entirely as local limit results. Thus the above argument will apply in each case. For brevity, we make no further mention of these technical details.

Now it is straightforward, using the density function and Stirling's formula, to show that for, say $W \sim$ Hypergeometric (N, M, n), with $M/N \to r$ and $n/N \to s$, then

$$\sqrt{n}\left(\frac{W}{n} - \frac{M}{N}\right) \xrightarrow{\mathrm{L}} \mathrm{N}(0, r(1-r)(1-s)).$$

Letting $W = D'$, $\mathrm{N} = \tilde k x_{\mathrm{n}}$, $M = x_{\mathrm{fn}}$, and $n = km$, (8), (15) and the above imply

$$\frac{x_{\mathrm{fn}}}{\tilde k x_{\mathrm{n}}} = \frac{x_{\mathrm{fn}}/\tilde N}{\tilde k x_{\mathrm{n}}/\tilde N} \xrightarrow{\mathrm{p}} \frac{\tilde k \pi(1-\eta)}{\tilde k q} \equiv \pi_1, \qquad \frac{mk}{\tilde k x_{\mathrm{n}}} \xrightarrow{\mathrm{p}} f,$$

$$\sqrt{km}\left(\frac{D'}{km} - \frac{x_{\mathrm{fn}}}{\tilde k x_{\mathrm{n}}}\right)\Bigg|_{F_{\tilde N}^{*}, V_{\tilde N}^{*}} \xrightarrow{\mathrm{L}} \mathrm{N}(0, \pi_1(1-\pi_1)(1-f))$$

and hence

$$\sqrt{\tilde N}\left(\frac{D'}{km} - \frac{x_{\mathrm{fn}}}{\tilde k x_{\mathrm{n}}}\right)\Bigg|_{F_{\tilde N}^{*}, V_{\tilde N}^{*}} \xrightarrow{\mathrm{L}} \mathrm{N}\left(0, \frac{\pi_1(1-\pi_1)(1-f)}{\tilde k q f}\right) \tag{16}$$

as $\tilde N \to \infty$. Thus (15) and (16) imply

$$\sqrt{\tilde N}\begin{pmatrix} \frac{mk}{\tilde k x_{\mathrm{n}}} - f \\ \frac{D'}{mk} - \frac{x_{\mathrm{fn}}}{\tilde k x_{\mathrm{n}}} \end{pmatrix}\Bigg|_{F_{\tilde N}^{*}} \xrightarrow{\mathrm{L}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_B \equiv \begin{pmatrix} \frac{f(1-f)}{\tilde k q} & 0 \\ 0 & \frac{\pi_1(1-\pi_1)(1-f)}{\tilde k q f} \end{pmatrix}\right), \tag{17}$$

since the limiting distribution in (16) is free from the limiting value V^{*}. Furthermore, since the limiting normal distribution in (17) is free of F^{*}, the limiting pair above is independent of the limiting random vector F^{*}.

Now consider $W_{\tilde N} \equiv (F_{\tilde N}'/\tilde N, mk/\tilde k x_{\mathrm{n}}, D'/mk - x_{\mathrm{fn}}/\tilde k x_{\mathrm{n}})$. Define $\gamma = (\mu_{\tilde k}', f, 0)'$. Then due to (8) and (17), we have established that, asymptotically,

$$\sqrt{\tilde N}(W_{\tilde N} - \gamma) \xrightarrow{\mathrm{L}} N_6(0, \text{Block Diag}\{\Sigma_F, \Sigma_B\}).$$

Then define $Y_{\tilde N} = (F_{\tilde N}', m/\tilde N, D'/\tilde N)'$, and the transformation $h(W) = (W_1, W_2, W_3, W_4, \tilde k W_3 W_5/k, \tilde k W_3 W_5 W_6 + W_4 W_5)'$. Then $h(W_{\tilde N}) = Y_{\tilde N}$, and

$$h(\gamma) = \begin{pmatrix} \tilde k \mu_1 \\ \mu_2 \\ \mu_3 \\ \tilde k \mu_4 \\ \tilde k f \mu_3/k \\ \tilde k f \mu_4 \end{pmatrix},$$

$$\dot h(\gamma) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & f\tilde k/k & 0 & \tilde k \mu_3/k & 0 \\ 0 & 0 & 0 & f & \tilde k \mu_4 & \tilde k \mu_3 f \end{pmatrix} \equiv \dot h(\gamma) = \begin{pmatrix} I_4 & 0 \\ G & H \end{pmatrix}.$$

It follows from the delta method that, asymptotically,

$$Y_{\tilde N}^{*} \equiv \sqrt{\tilde N}(Y_{\tilde N} - h(\gamma)) \xrightarrow{\mathrm{L}} N_6(0, \Sigma_Y \equiv \dot h(\gamma) \text{ Block Diag } \{\Sigma_F, \Sigma_B\} \dot h(\gamma)'), \tag{18}$$

where

$$\Sigma_Y = \begin{pmatrix} \Sigma_F & \Sigma_F G' \\ G\Sigma_F & G\Sigma_F G' + H\Sigma_B H' \end{pmatrix},$$

$$G\Sigma_F = \tilde k f \begin{pmatrix} -\tilde k \mu_1 \mu_3/k & -\mu_2\mu_3/k & \mu_3(1-\mu_3)/k & \tilde k \mu_4(1-\mu_3)/k \\ -\tilde k \mu_1 \mu_4 & -\mu_2\mu_4 & \mu_4(1-\mu_3) & \mu_4\{1 - \tilde k \mu_4 + \pi(\tilde k - 1)\} \end{pmatrix}.$$

$$G\Sigma_F G' + H\Sigma_B H'$$
$$= \tilde k f \begin{pmatrix} \mu_3\{1 - f + \tilde k f(1-\mu_3)\}/k^2 & \mu_4\{1 - f + \tilde k f(1-\mu_3)\}/k \\ * & \mu_4\{1 - f\tilde k \mu_4 + f\pi(\tilde k - 1)\} \end{pmatrix}.$$

In order to obtain the limiting distribution of $s_{\mathrm{tp}}^{\mathrm{g}}$, we require the joint limiting distribution of the vector of counts, say $\{D_j\colon j = 1, \ldots, k\}$, where D_j is the number of second-stage groups with exactly j D's. It is straightforward to show that

$$\mathrm{pr}(\{D_j\colon j=1,\ldots,k\}|F_{\bar N},D',m)=\frac{\prod_{j=1}^{k}\binom{k}{j}^{D_j}\binom{m}{\{D_j\}}}{\binom{mk}{D'}}. \tag{19}$$

Define normalized values for the D_j's, and for D',

$$D^*_{mj}=\sqrt{m}\left(\frac{D_j}{m}-p_j\right),$$

$$p_j=\binom{k}{j}\pi_1^j(1-\pi_1)^{k-j},\quad j=1,\ldots,k,\ D'^*_m=\sqrt{km}\left(\frac{D'}{km}-\pi_1\right).$$

Note that $\sum_{j=1}^{k} jD^*_{mj}=\sqrt{k}D'^*_m$ since $\sum jD_j=D'$.

In Section 5.3, and using (19), we establish conditional joint asymptotic normality of the D^*_{mj}'s by proving a local limit theorem. Defining $Z^*_m=\{D^*_{mj}\colon j=1,\ldots,k\}$, it is thus established that

$$Z^*_m|_{F^*_{\bar N},D'^*_m,m}\xrightarrow{\mathrm{L}}Z^*\sim \mathrm{RN}\left(0,\Sigma_D;\sum_{j=1}^{k}jZ^*_j=\sqrt{k}D'^*\right),$$
$$\Sigma_D=\mathrm{Diag}\{p\}-pp',\qquad p=(p_1,\ldots,p_k)' \tag{20}$$

as $m\to\infty$, and where $\mathrm{RN}(\cdot,\cdot,\cdot)$ denotes a restricted (singular) normal distribution with given mean vector, covariance, and restriction, and where we have conditioned on the sequence of possible values D'^*_m such that their limit is D'^*. Since the distribution is free from F^*, F^* and Z^* are independent, conditional on D'^*.

Remark. As a check, we should be able to obtain the distribution of the limiting random variable $D'*=\sum jZ^*_j/\sqrt{k}$ from the unrestricted distribution for Z^*. Since $\tilde{a}'\Sigma_D\tilde{a}/k=\pi_1(1-\pi_1)$, this is indeed the case.

Finally, we obtain the limiting conditional distribution of $s^{\mathrm{g}}_{\mathrm{tp}}$. Let $x^{\mathrm{g}}_{\mathrm{tp}j}$ denote the number of second-stage true positive groups that contain exactly j D's. Then $s^{\mathrm{g}}_{\mathrm{tp}}=\sum_j jx^{\mathrm{g}}_{\mathrm{tp}j}$ and $x^{\mathrm{g}}_{\mathrm{tp}}=\sum_j x^{\mathrm{g}}_{\mathrm{tp}j}$. Conditional on the values $(\{D_j\},mD')$, $x^{\mathrm{g}}_{\mathrm{tp}j}$'s are independent and Binomial, namely

$$x^{\mathrm{g}}_{\mathrm{tp}j}|_{F_{\bar N},D',\{D_j\},m}\sim \mathrm{Bin}(D_j,\eta_{\mathrm{g}}).$$

Thus if the D_j's are known, the $x^{\mathrm{g}}_{\mathrm{tp}j}$'s are independent of everything else. Since $D_j\to\infty$ in probability as $m\to\infty$, it follows that the $\sqrt{D_j}(x^{\mathrm{g}}_{\mathrm{tp}j}/D_j-\eta_{\mathrm{g}})$ converge, conditionally, to independent $\mathrm{N}(0,\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}}))$ random variables. Furthermore, since $D_j/m\to p_j$ in probability as $m\to\infty$, we obtain

$$V^*_{mj}\equiv\sqrt{m}(x^{\mathrm{g}}_{\mathrm{tp}j}/D_j-\eta_{\mathrm{g}})\xrightarrow{\mathrm{L}}V^*_j\sim \mathrm{N}(0,\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})/p_j)$$

independently as $m\to\infty$. Since this limiting distribution is free from (F^*,D'^*,Z^*), we have asymptotic independence. Then defining $V_j=x^{\mathrm{g}}_{\mathrm{tp}j}/D_j$, we have established

$$V^*_m=\{\sqrt{m}(V_j-\eta_{\mathrm{g}})\colon\ j=1,\ldots,k\}\xrightarrow{\mathrm{L}}V^*\sim N_k(0,\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})\mathrm{Diag}\{1/p\}). \tag{21}$$

Finally, we require the limiting joint distribution of $T^*_m\equiv\{T^*_{mj}\equiv\sqrt{m}(x^{\mathrm{g}}_{\mathrm{tp}j}/m-p_j\eta_{\mathrm{g}})\colon j=1,\ldots,k\}$. Defining $Z_j=D_j/m$, we note that $T^*_{mj}=\sqrt{m}(V_jZ_j-p_j\eta_{\mathrm{g}})$. The limiting distribution we require is then the joint distribution of $T^*_m=\{\sqrt{m}(Z_jV_j-p_j\eta_{\mathrm{g}})\colon j=1,\ldots,k\}$. Define the transformation $g(Z,V)=\{Z_jV_j\colon j=1,\ldots,k\}$. Define $\dot{g}(\cdot,\cdot)$ to be the derivative matrix of the transformation. Then $\dot{g}(Z,V)=(\mathrm{Diag}\{V\},\mathrm{Diag}\{Z\})$. Evaluating, we obtain $\dot{g}(p,\eta_{\mathrm{g}}e_k)=(\eta_{\mathrm{g}}I_k,\mathrm{Diag}\{p\})$, and hence using (20), (21) and the delta method,

$$T^*_m\xrightarrow{\mathrm{L}}\eta_{\mathrm{g}}Z^*+\mathrm{Diag}\{p\}V^*,,\ \ Z^*\perp V^*. \tag{22}$$

Thus normalizing $s^{\mathrm{g}}_{\mathrm{tp}}=\sum jx^{\mathrm{g}}_{\mathrm{tp}j}$, we obtain, conditionally

$$\begin{aligned}\sqrt{m}\left(\frac{s^{\mathrm{g}}_{\mathrm{tp}}}{m}-k\pi_1\eta_{\mathrm{g}}\right)\Bigg|_{F^*_{\bar N},D'^*_m,m}&\xrightarrow{\mathrm{L}}\tilde{a}'(\eta_{\mathrm{g}}Z^*+\mathrm{Diag}\{p\}V^*)\\&\sim\textstyle\sum jZ^*_j\eta_{\mathrm{g}}+\mathrm{N}\left(0,\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})\sum j^2p_j\right)\\&=\mathrm{N}(\sqrt{k}D'^*\eta_{\mathrm{g}},\,k\pi_1\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})\{1+\pi_1(k-1)\}).\end{aligned} \tag{23}$$

Similarly, and due to (20), we have unconditionally

$$\begin{pmatrix}\sum_{j=1}^{k}Z^*_j\\ \sum_{j=1}^{k}jZ^*_j\end{pmatrix}=\begin{pmatrix}e'_k\\ \tilde{a}'\end{pmatrix}Z^*\sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}p_0(1-p_0) & k\pi_1p_0\\ k\pi_1p_0 & k\pi_1(1-\pi_1)\end{pmatrix}\right)$$

and using this and (22), we obtain, conditionally,

$$\sqrt{m}\left(\frac{x^{\mathrm{g}}_{\mathrm{tp}}}{m}-(1-p_0)\eta_{\mathrm{g}}\right)\xrightarrow{\mathrm{L}}e'_k(\eta_{\mathrm{g}}Z^*+\mathrm{Diag}\{p\}V^*)\sim\eta_{\mathrm{g}}\textstyle\sum Z^*_j|\sum jZ^*_j \tag{24}$$

$$\begin{aligned}&=\sqrt{k}D'^*+\mathrm{N}(0,\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})(1-p_0))\\&=\mathrm{N}\left(\eta_{\mathrm{g}}\sqrt{k}D'^*\frac{p_0}{1-\pi_1},\eta_{\mathrm{g}}^2\left\{p_0(1-p_0)-\frac{k\pi_1p_0^2}{1-\pi_1}\right\}\right.\\&\qquad\left.+\,\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})(1-p_0)\right).\end{aligned} \tag{24}$$

A.2. Joint asymptotic distribution of $(x_{\mathrm{tp}},x_{\mathrm{fp}},x_{\mathrm{n}},s^{\mathrm{g}}_{\mathrm{tp}})$

By (23), we have established

$$\begin{aligned}\sqrt{m}\left(\frac{s^{\mathrm{g}}_{\mathrm{tp}}}{m}-\frac{D'\eta_{\mathrm{g}}}{m}\right)\Bigg|_{F^*_{\bar N},D'^*_m,m}&=\left\{\sqrt{m}\left(\frac{s^{\mathrm{g}}_{\mathrm{tp}}}{m}-k\pi_1\eta_{\mathrm{g}}\right)\right.\\&\qquad\left.-\eta_{\mathrm{g}}\sqrt{k}\sqrt{mk}\left(\frac{D'}{mk}-\pi_1\right)\right\}\Bigg|_{F^*_{\bar N},D'^*_m,m}\\&\xrightarrow{\mathrm{L}}N_1(0,k\pi_1\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})\{1+\pi_1(k-1)\})\end{aligned} \tag{25}$$

which is free from all conditioning. Since $m/\tilde{N} \to \tilde{k}fq/k$ in probability as $\tilde{N} \to \infty$, we define $m^*_{\tilde{N}} = \sqrt{\tilde{N}}(m/\tilde{N} - \tilde{k}fq/k)$. Defining $D'^*_{\tilde{N}} = \sqrt{\tilde{N}}(D'/\tilde{N} - \tilde{k}f\mu_4)$, we obtain from (25) that,

$$\sqrt{\tilde{N}}\left(\frac{s^{\mathrm{g}}_{\mathrm{tp}}}{\tilde{N}} - \frac{D'\eta_{\mathrm{g}}}{\tilde{N}}\right)\Bigg|_{F^*_{\tilde{N}}, D'^*_{\tilde{N}}, m^*_{\tilde{N}}} \xrightarrow{\mathrm{L}} N_1(0, \sigma^2_b \doteq \tilde{k}fq\pi_1\eta_{\mathrm{g}}(1-\eta_{\mathrm{g}})\{1+\pi_1(k-1)\}). \tag{26}$$

We are now in a position to obtain the joint asymptotic distribution of $(x_{\mathrm{tp}}, x_{\mathrm{fp}}, x_{\mathrm{n}}, s^{\mathrm{g}}_{\mathrm{tp}})$. Recall that the asymptotic distribution of normalized $Y_{\tilde{N}} = \{(x_{\mathrm{tp}}/\tilde{N}, x_{\mathrm{fp}}/\tilde{N}, x_{\mathrm{n}}/\tilde{N}, x_{\mathrm{fn}}/\tilde{N}, m/\tilde{N}, D'/\tilde{N})'$, namely $Y^*_{\tilde{N}}$, is obtained from (18) as $N_6(0, \Sigma_Y)$. Then using (18) and (26) we have, unconditionally,

$$\begin{pmatrix} Y^*_{\tilde{N}} \\ \sqrt{\tilde{N}}(s^{\mathrm{g}}_{\mathrm{tp}}/\tilde{N} - D'\eta_{\mathrm{g}}/\tilde{N}) \end{pmatrix} \xrightarrow{\mathrm{L}} N_7(0, \text{Block Diag}\{\Sigma_Y, \sigma^2_b\}).$$

Define $Y^+_{\tilde{N}} = (Y'_{\tilde{N}}, s^{\mathrm{g}}_{\mathrm{tp}}/\tilde{N})', \gamma^+ = (h(\gamma)', \tilde{k}f\eta_{\mathrm{g}}\mu_4)'$ and $Y^{+*}_{\tilde{N}} = \sqrt{\tilde{N}}(Y^+_{\tilde{N}} - \gamma^+)$. Then another application of the delta method yields

$$Y^{+*}_{\tilde{N}} \xrightarrow{\mathrm{L}} N_7(0, \Sigma_{Y^+}), \quad \Sigma_{Y^+} \equiv \begin{pmatrix} \Sigma_Y & d\eta_{\mathrm{g}} \\ d'\eta_{\mathrm{g}} & \sigma^2_b + \eta^2_{\mathrm{g}}\sigma^2_e, \end{pmatrix},$$

where d is the submatrix of Σ_Y, defined at (18), corresponding to the asymptotic covariances for normalized D', and σ^2_e is the corresponding asymptotic variance. The transformation was of the form $s(w) = (w_1, \ldots, w_6, w_7 + \eta_{\mathrm{g}}w_6)'$. Finally, after some algebra and integrating out unobservables, we obtain the joint asymptotic marginal distribution for $(x_{\mathrm{tp}}, x_{\mathrm{fp}}, x_{\mathrm{n}}, s^{\mathrm{g}}_{\mathrm{tp}})$, namely result (10).

A.3. *Proof of* (20)

Starting with (19) and using Stirling's approximation, we obtain

$$\mathrm{pr}(\{D_j\colon j=1,\ldots,k\}|F_{\tilde{N}}, D'_m, m)$$

$$= \frac{\prod_{j=1}^k \binom{k}{j}^{D_j} (2\pi m)^{-(k-1)/2}\sqrt{k}(D'/mk)^{D'+1/2}(1-D'/mk)^{mk-D'+1/2}}{\prod_{j=1}^k (D_j/m)^{D_j+1/2}(1-D_+/m)^{m-D_+ +1/2}}$$

$$\times(1+\mathrm{O}(m^{-1})), \tag{27}$$

where $D_+ = \sum_{j=1}^k D_{mj}$, the number of second-stage groups with at least one defective. Recall the definitions just after (16) and furthermore define $D^*_{m+} = \sum_{j=1}^k D^*_j$. Then

$$\frac{D_j}{m} \doteq p_j + \frac{D^*_{mj}}{\sqrt{m}}, \qquad 1 - \frac{D_+}{m} = p_0 - \frac{D^*_{m+}}{\sqrt{m}}, \qquad \sum_{j=1}^k jD^*_{mj} = \sqrt{k}D'^*_m,$$
$$\frac{D'}{mk} = \pi_1 + \frac{D'^*_m}{\sqrt{km}}. \tag{28}$$

Substituting (28) into (27), and simplifying we obtain

$$\mathrm{pr}(\{D_j\colon j=1,\ldots,k\}|F^*_{\tilde{N}}, D'^*_m, m)$$
$$= \prod_{j=1}^k \binom{k}{j}^{-1/2} (2\pi m)^{-(k-1)/2}\sqrt{k}\{\pi_1(1-\pi_1)\}^{-(k(k+1)/4)+(1/2)}B_m(1+\mathrm{O}(m^{-1})). \tag{29}$$

$$B_m = \frac{(1+D'^*_m/\sqrt{km}\pi_1)^{D'}(1-D'^*_m/\sqrt{km}(1-\pi_1))^{mk-D'}}{\prod_{j=1}^k (1+D^*_{mj}/\sqrt{m}p_j)^{D_j}(1-D^*_{m+}/\sqrt{m}p_0)^{m-D_+}}.$$

Furthermore,

$$\ell n(B_m) = -\frac{1}{2}\left\{\sum_{j=1}^k (D^*_{mj})^2/p_j + (D^*_{m+})^2 - (D'^*_m)^2/(\pi_1(1-\pi_1))\right\} + \mathrm{O}((m)^{-1/2})$$

$$= -\frac{1}{2}\left\{\sum_{j=1}^k (D^*_{mj})^2\left(\frac{1}{p_j}+\frac{1}{p_0}\right) + 2\sum_{j<j'} D^*_{mj}D^*_{mj'}/p_0 - (D'^*_m)^2/(\pi_1(1-\pi_1))\right\}$$

$$+\mathrm{O}((m)^{-1/2}).$$

Define the $k \times k$ matrix $C = c_{ij}$ where $c_{ii} = (1/p_j + 1/p_0)$ and $c_{ij} = 1/p_0$. Then $C = \mathrm{Diag}\{1/p\} + e_ke'_k/p_0$ and $C^{-1} = \mathrm{Diag}\{p\} - pp'$ where $p \equiv (p_1, p_2, \ldots, p_k)'$. Recall that $\Sigma_D = C^{-1}$ and $Z^*_m = (D^*_{m1}, \ldots, D^*_{mk})'$. Then

$$\ell n(B_m) = -\tfrac{1}{2}\{(Z^*_m)'\Sigma_D^{-1}Z^*_m - (D'^*_m)^2/(\pi_1(1-\pi_1))\} + \mathrm{O}((m)^{-1/2}). \tag{30}$$

We furthermore note that

$$|\Sigma_D| = |\mathrm{Diag}\{p\}||I_k - \mathrm{Diag}\{1/p\}pp'| = \prod_{j=1}^k p_j(1 - p'\mathrm{Diag}\{1/p\}p)$$

$$= \left\{\prod_{j=1}^k p_j\right\} p_0 = \prod_{j=0}^k \binom{k}{j} \pi_1^j(1-\pi_1)^{k-j}$$

$$= \prod_{j=0}^k \binom{k}{j} \pi_1^{k(k+1)/2}(1-\pi_1)^{k(k+1)/2}. \tag{31}$$

So consider the lattice

$$\left\{Z^*_{mj} = \sqrt{m}(D_j/m - p_j)\colon j=1,\ldots,k, \textstyle\sum jZ^*_{mj} = \sqrt{k}D'^*_m, \text{ all possible } D_j \text{ and } m\right\}.$$

The volume of a rectangle defined on this lattice is $m^{(k-1)/2}$. We have thus established the local limit theorem (Okamoto, 1959) by utilizing (30) and (31) in (29), and by choosing a sequence of values from the above lattice such that they converge to

$: 1,\ldots,k\} = Z^*$, namely

$$\mathrm{im}_{\to\infty}\ m^{(k-1)/2}\mathrm{pr}(\{D_j\}|F_N^*, D_m'^*, m) = \frac{n_k(Z^*;0,\Sigma_D)}{N_1(\sum jZ_j^*;0,k\pi_1(1-\pi_1))}, \qquad (32)$$

of a $N_k(0,\Sigma_D)$ p.d.f and a $\mathrm{N}(0,k\pi_1(1-\pi_1))$ p.d.f and where the Z_j^*'s are
1 so that $\sum_{j=1}^k jZ_j^* = \sqrt{k}D'^*$. We have now established (20) since (32) implies
nce in law.

ces

oung, B., et al., 1989. Sensitivity and specificity of pooled vs. individual testing in HIV antibody
nce study. J. Clin. Microbiol. 27, 1893–1895.
., Swallow, W.H., 1990. Using group testing to estimate a proportion, and to test the binomial
Biometrics 46, 1035–1046.
., 1996. Bayesian models for limiting dilution assay and group test data. Biometrics 52, 1055–1062.
R., 1943. The detection of defective members of large populations. Ann. Math. Statist. 14,
10.
I, J.C., Bassett, M.T., Smith, H.J., Jacobs, J.A., 1988. Pooling of sera for human immunodeficiency
HIV) testing: An economic method for use in developing countries. J. Clin. Pathol. 41, 582–585.
A.A., 1998. Advances in HIV/AIDS statistical methodology over the past decade. Statist. Med. 17,

J.L., 1987. The statistical precision of medical testing procedures: application to polygraph/AIDS
ies test data. Statist. Sci. 2, 213–238.
J.L., Hammick, P.A., 1989. Estimation of the prevalence of rare disease preserving the anonymity
subjects by group testing: application to estimating the prevalence of AIDS anitbodies in blood
. J. Statist. Plann. Inference 22, 15–27.
. J.L., Johnson, W.O., Reneau, D.M., 1991. Bayesian analysis of screening data: applications to
in blood donors. Can. J. Statist. 19, 135–150.
. J.L., Johnson, W.O., 1994. Screening with cost-effective quality control: potential applications to
id drug testing. J. Amer. Statist. Assoc. 89, 972–981.
I.H., Rawlinson, V.I., 1988. HIV antibody screening of blood donations in the United Kingdom.
ang 54, 34–38.
, P.A., Gastwirth, J.L., 1994. Group testing for sensitive characteristics: extension to higher
nce levels. Internat. Statist. Rev. 62, 319–331.
liver, J.M., Swallow, W.H., 1994. A two-stage adaptive group-testing procedure for estimating
proportions. J. Amer. Statist. Assoc. 89, 982–993.
W.O., Gastwirth, J.L., 1991. Bayesian inference for medical screening tests: approximations useful
analysis of acquired immune deficiency syndrome. J. Roy. Statist. Soc. B 53, 427–439.
. M.L., Koskela, P., Leinikki, P., 1996. Unlinked anonymous HIV screening of pregnant women in
prevalence population. Scand. J. Infectious Dis. 28, 3–7.
L., Frother, T.A., Brookmeyer, R. et al., 1989. Evaluation of human immunodeficiency virus
valence in population surveys using pooled sera. J. Clin., Microbiol. 27, 1449–1452.
Bounlu, K., Pholsena, V., Mastro, T.D., 1998. HIV seroprevalence among pregnant women in
ne, Laos. AIDS 12, 1403.
., Tu, X.M., Pagano, M., 1994. Screening for the presence of disease by pooling sera samples.
r. Statist. Assoc. 89, 424–434.
r, J., Chiavetta, J., Naiman, R., Buchner, B., Scalia, V., Horst, R., 1986. Evaluation of a confidential
I of excluding blood donors exposed to human immunodeficiency virus. Transfusion 26, 539–541.
O.T., Paladin, F.J.E., Dimaandal, E., Balis, A.M., Samson, C., Mitchell, S., 1991. Relevance of
ly content and test format in HIV testing of pooled sera. AIDS 6, 43–47.
M.A., 1959. A convergence theorem for discrete probability distributions. Ann. Inst. Statist. Math.
7–112.

Ratcliffe, J., Ades, A., Gibb, D., Sculpher, M.J., Briggs, A.H., 1998. Prevention of mother to child transmission of HIV-1 infection: alternative strategies and their cost-effectiveness. AIDS 12, 1381–1388.

Sethuraman, J., 1961. Some limit theorems for joint distributions. Sankhya, series A 23, 379–386.

Sherlock, C.H., Strathdee, S.A., Le, Tom, Sutherland, D., O'Shaughnessy, M.V., 1995. Use of pooling and outpatient laboratory specimens in an anonymous seroprevalence survey of HIV infection in British Columbia, Canada. AIDS 9, 945–950.

Thompson, K.H., 1962. Estimation of the proportion of vectors in a natural population. Biometrics 18, 568–578.